DOCUMENT RESUME

ED 333 036 TM 01C 545

AUTHOR Moore, Mary Ann
TITLE The Place of Significance Testing in Contemporary Social Science.
PUB DATE 3 Apr 91
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Data Interpretation; Educational Research; *Effect Size; Research Methodology; *Research Problems; *Sample Size; *Social Science Research; *Statistical Significance

ABSTRACT

This paper examines the problems caused by relying solely on statistical significance tests to interpret results in contemporary social science. The place of significance testing in educational research has often been debated. Among the problems in reporting statistical significance are questions of definition and terminology. Problems are also found in the use, as well as the reporting, of significance testing. One of the most important problems is the effect of sample size on significance. An example with a fixed effect size of 25% and samples containing 22, 23, and 24 people illustrates these effects. The issues of validity and reliability in significance testing with measurement studies are considered. Although these problems are widely recognized, publishers show a clear bias in favor of reports that claim statistical significance. Researchers need to recognize the limitations of significance testing. Effect size statistics aid in the interpretation of results and provide a guide to the relative importance of the study. Two tables illustrate the effects of sample size. A 22-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED330036

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MARY ANN MOORE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Place of Significance Testing in Contemporary Social Science

Mary Ann Moore

Texas A&M University 77843-4225

Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 3, 1991.

BEST COPY AVAILABLE

TM016545

ABSTRACT

The place of significance testing in educational research has been often debated (Cohen, 1990; Hays, 1981; Rosnow & Rosenthal, 1989; Thompson, 1989a). This paper examines the problems caused by relying solely on statistical significance tests to interpret results in contemporary social science. An example is used to illustrate the effect of sample size on significance tests. Ways of evaluating this influence to interpret results are explored using heuristic examples.

The Place of Significance Testing in Contemporary Social Science

Educational researchers have historically relied on statistical significance tests to evaluate the results of their research studies. The use of significance testing in this manner has been much debated (Chow, 1988; Shaver, 1979; Thompson, 1989a). Specifically, the effect of sample size on significance testing has been well documented (Hays, 1981; Meehl, 1978; Rosnow & Rosenthal, 1989). This paper explores these issues as well as some potential improvements in practice. Additionally, questions regarding the applicabilitity of significance testing to measurement studies is discussed. Lastly, the role of effect size statistics and their possible implications for reporting results will be examined.

Trouble with Reporting Statistical Significance

There are three main problems in using the term 'significant' to describe research results. First, many times the word 'statistically' is not used in the description of the results. Researchers may report that a 'significant diference was obtained'. This statement does imply the word 'statistically,' but many people may not understand the implication or hint regarding what is being discussed.

The second problem using the term 'statistically significant' focuses on the second part of the description,

'significant'. What does it mean when one uses the word 'significant'? Webster's Dictionary defines significance as importance, weight, or meaning. Is it possible to correctly replace one of those terms with 'significant' in the description of results that are found to be 'statistically significant'? It is a misnomer to describe results as 'statistically important' or 'statistically meaningful'. As Holmes (1990, p. 72) notes, "The fact that something is statistically significant in no way means the result is of any importance or has any real-life significance" (Holmes, 1990, p.72).

The last problem in using 'statistical significance' to describe results is that it is often used with a qualifier to explain its use. For example, one statement that is often found in the literature is 'The results approached significance'. This statement implies that the data is closer to reaching significance than it is away from it. This is simply not true, but by making such a statement the researcher may hope that the naive reader believes that the study is 'important' or has 'meaning.' It is equally plausible (and just as anthropromorphic) to interpret nonsignificant results as results that were 'trying to avoid' being statistically significant!

To illustrate the difference between stating a result that is 'significant' or 'important' a hypothetical study using I.Q. scores will be used. The purpose of the study was to compare

intelligence levels of high school students from the private high schools to the public high schools. The mean I.Q. (100.15, SD=15) of the 12,000 private high school students was significantly significant (Zcalc=2.12>Zcrit=1.96, p<.05) from the mean (99.85, SD=15) of the 188,000 public high school students in the public schools. Is it possible to conclude that the brighter students choose private high schools over public high schools? Or is the education superior at private vs. public high schools? The naive researcher may jump to some of these conclusions, but these interpretations do not accurately describe the results.

It is important to recognize that the standardized difference between the two means (.3/15=.002) is meaningless. The difference of .3 or one-third of an I.Q. point for all practical purposes means nothing. Based on this study, there is no reason to expect that there are meaningful differences between private and public high school student's I.Q. scores, even though the difference is 'significant.'.

Trouble With Using Significance Testing

The role of using significance testing in research has been often in debate in the literature (Chow, 1988; Shaver, 1979; Thompson, 1989a). An extreme view of this debate was taken by Carver (1978). He recommended that all statistical significance testing in educational research should be

abandoned. Carver (1978) stated three reasons why statistical significance testing, by itself, means nothing. The first reason he calls the odds-against-chance fantasy. The following example will help illustrate this point. The probability of a person being dead, given they were hanged, may be .97 or higher, p(D/H). The probability of being hanged, given that the person is dead, may be .01 or lower, p(H/D). If one attempted to reverse the question it does not always work. Yet this is the logical that is being used when researchers obtain statistically significant results and then state that the null hypothesis can be rejected and therefore the study must support the alternative hypothesis.

The second reason is called replicability or reliability fantasy.

> Whether or not the results are replicable or reliable actually depends upon whether the important variables can be controlled and manipulated in exactly the same way in order to give the same results. (Carver, 1978, p. 387)

Carver emphasizes that researchers too often substitute statistical significance testing for actual replicative evidence.

The third reason why statistical significance means nothing according to Carver (1978) involves the valid research hypothesis fantasy. He states that, "Even if the null hypothesis

can be rejected, several other alternatives or rival hypotheses still must be ruled out" (Carver, 1978, p. 387).

Huberty (1987) partially supports Carver's view of the problems of statistical significance testing, yet does not support altogether abandoning such testing. Huberty argued that statistical tests should be used as guides and indicators, not as a means of arriving at definite answers. Likewise, Rosnow and Rosenthal (1989) stressed the importance of significant testing in publications, however they clearly do not support basing decisions of publication completely on significance testing.

> It may not be an exaggeration to say that for many Ph.D. students, for whom the .05 alpha has acquired an almost ontological mystique, it can mean joy, a doctoral degree, and a tenure-track position at a major university if their dissertation p is less than .05 . . . [But] surely, God loves the .06 nearly as much as the .05 [level]. (Rosnow & Rosenthal, 1989, p. 1277)

Based on this brief discussion of the cautions and limitations of using statistical significance testing in educational research, it is apparent that there are problems associated with making decisions solely based on such testing. One major issue when using significance testing is the role of the sample size.

<u>The Effect of Sample Size on Significance</u>

It has been argued that "virtually any study can be made to show significant results if one uses enough subject" (Hays, 1981, p. 293). Meehl (1978) also stressed that if researchers take the null hypothesis literally, it is always false. Recent research has begun to recognize the limitations of significance testing (Kupfersmid, 1988; Rosnow & Rosenthal, 1989). As Thompson (1989b, p. 66) notes, "Because statistical significance is largely an artifact of sample size, significance decisions must be interpreted in the context of sample size." The following example will be used to illustrate the influence of sample size on significance test results.

Presume that a researcher conducted a one-way ANOVA where there were two levels of an independent variable. The sum-of-squares total was 100, and one independent variable explained a quarter of the variation in the dependent variable (i.e., eta-squared or correlation ratio equaled .25 or 25%).

Table 1 illustrates when a nonsignificant result becomes significant. The table presents significance tests where different sample sizes and fixed effect sizes (25%) are used.

INSERT TABLE 1 ABOUT HERE

For the fixed effect size (25%), with a sample of 22 people, this result is not statistically significant. However, for

the 25% effect size reported in Table 1, the result changes from not being statistically significant with a sample of 23 ($F=3.33$, $p>.05$) to reaching significance using 24 as a sample ($F=3.5$, $p<.05$). It is important to keep in mind that the effect size did not change.

Table 2 illustrates that sample sizes have a similar effect on significance testing when using more than two levels of independent variables. Table 2 uses three levels of an independent variable and demonstrates similar results.

INSERT TABLE 2 ABOUT HERE

For the fixed 25% effect size fixed, the decision to reject or not reject the null hypothesis can be changed by adding or dropping a single subject. Using 30 subjects yields a nonsignificant result ($F=2.89$, $p>.05$), yet with 31 subjects the null hypothesis is rejected ($F=3.0$, $p<.05$). This further illustrates the role of sample size on significance test results.

The purpose of this illustration was not to teach the reader how many subjects are necessary to reach statistical significance or how many subjects are necessary to reach significance given an increase in the levels of the independent variables. Rather, the purpose was to provide the reader with

a reasonable context in which to evaluate significant test results (Thompson, 1988).

Bias Against Nonsignificance in Publications

Based on the previous discussion of the effect of sample size on significance testing, it is difficult to understand the importance of obtaining significant results in order for a study to be published. But as Thompson (1987, p. 11) notes, "There can be little question but that the published literature historically reflects a major bias against statistically non-significant results.". Atkinson, Furlong and Wampold (1982) examined the importance of statistical significance in publishing. The authors asked consulting editors of two major psychological publications (Journal of Counseling Psychology and Journal of Consulting and Clinical Psychology) to evaluate three versions of a research manuscript that differed only with respect to the level of statistical significance. They reported that "The statistically nonsignificant and approach significance versions were more than three times as likely to be recommended for rejection than was the statistically significant version" (Atkinson et al., 1982, p. 189).

There are other problems in stressing publishing studies that yield significant results, in addition to the previous mention problem of the effect of sample size. One major problem exists when researchers not only fail to even submit a

manuscript that contains nonsignificant results but also give up the entire direction of their research (Greenwald, 1975).

Another problem exists in the education of researchers about the importance of significance testing. It is stressed by the majority of educational statistics professors that students need to reach statistical significance and that in fact is the goal of research. The last problem focuses on the justification of finding significant or nonsignificant results. When a researcher fails to replicate an earlier study explanations are generally made in reference to sampling error or problems with the sample population. Yet, why can't an initial study be the result of these same problems and the same influences be an explanation for nonsignificant replication results? (Atkinson et al., 1982).

Clearly, there are problems in relying solely on significance testing to determine the possibility of publication or the importance of a study's results. Significance testing was not developed for this purpose (Fisher, 1925), although in educational research it is often used in this manner. The apparent problems in using statistical significance testing is even greater in it's use with test and measurement studies.

Significant Testing With Measurement Studies

Validity

There are two types of criterion-related validation: predictive and concurrent. "Predictive validity refers to the degree to which test scores predict criterion measurements that will be made at some point in the future" (Crocker & Algina, 1986, p. 224). An example that is often used to illustrate predictive validity is the use of Scholastic Aptitude Test (SAT) scores in making college admissions decisions. For instance, SAT scores may correlate about .50 with college grade point average (GPA). As a result, SAT scores have a degree of predicted validity in relationship to GPA and therefore college administrators are at least partially justified in using SAT scores in admission decisions.

As Anastasi (1988, p. 146) explains, "Concurrent validation is relevant to tests employed for diagnosis of existing status, rather than prediction of future outcomes." One way to demonstrate the difference between these two types of validity is to examine whether a counselor wants to know: "Is Jones depressed?" (concurrent validation) or "Is Jones likely to become depressed?" (predictive validation).

In determining whether a result is statistically significant, historically researchers have set the null hypothesis at zero. In relation to validity studies this means that the null hypothesis states that there will be absolutely no relationship between the test instrument and the observable behavior. For

example, if mental patients took the Beck's Depression Inventory and immediately afterward were observed and rated on the depressed behaviors displayed on the ward, a positive correlation would be evidence of concurrent validity of the paper-and-pencil test. Yet, by giving patients in a mental hospital a test for depression, an assumption is being made that the patient is depressed and will exhibit behaviors that reflect depression. Testing a "no difference" null makes little sense in such contexts.

Reliability

An individual using a specific test needs to have some confidence that the results of the test could be replicated if the same people were to take the exact same test under similar circumstances. Simply stated, reliability refers to the desired consistency of test scores. As Crocker and Algina (1986, p. 105) explain, "In practical terms reliability is the degree to which individuals' deviation scores, or z-scores, remain relatively consistent over repeated administrations of the same test or alternative test forms.". Theoretically, a reliability coefficient is the correlation between scores on parallel test forms or mathematically defined as the ratio of true score variance to observed score variance. It is important to understand that this is a theoretical concept, for truly parallel tests do not occur in the practical sense. Since this is the case,

there have been developed many procedures used to estimate test score reliability. Three such procedures that require two test administrations are: alternative forms; test-retest; and test-retest with alternative forms. There are other methods of estimating reliability that require a single test administration such as: split half methods and methods based on item covariances.

When computing an estimate of the reliability coefficient the desired coefficient is greater than .80. One is to be wary of an instrument where the estimated reliability coefficient is less than .80. Therefore, in terms of reliability coefficients, once again statistical significance of a "no difference" null hypothesis does not make much sense.

The Role of Effect Size in Reporting Results

Effect size attempts to answer "Is the sample size large enough to give confidence that the big associations will indeed show up, while being small enough so that trivial associations will be excluded from significance?" (Hays, 1981, p. 294). Two arguments for the use of effect size estimates procedures have been postulated (Rosnow & Rosenthal, 1989). First, by computing population effect sizes this may help give the researcher guidance about the sample size needed for future studies. Second, and more critically, "it is important to realize that the effect size tells us something very different from the p

level" (Rosnow & Rosenthal, 1989, p.1274). Statistically significant results are not necessarily practically significant based on the size of the effect. Specifically, high significant p values do not automatically reflect large effects.

In regards to result importance a variety of effect size statistics have been developed. They range from simple to sophisticated in their computations. The simplest effect sizes statistics are analogous to the coefficient of determination (squared r), and is called eta-squared (Hays, 1981, p. 349). As Welge-Crow, LeCluyse and Thompson (1990, p. 6) note, "Such statistics inform the researcher regarding what proportion of variance in the dependent variable(s) is explained by a given predictor." A more complex and sophisticated effect size statistic is the corrected version of the multivariate omega squared statistic (Tatsuoka, 1973, p. 30), and other effect size estimates exist as well.

The question that needs to be asked is why more researchers don't utilize effect size statistics when reporting results. There does not appear to be one simple answer to this question. "The failure to utilize effect size estimates in interpretation may partially be due to the influence of the significance testing paradigm which influences attention away from effect size" (Thompson, 1987, p. 23). Another explanation may be that researchers fail to recognize that all analyses

actually test association, and therefore that effect sizes analogous to the coefficient of determination are appropriate for all analytic methods. Thompson (1989c) argues that the failure to see that all analytic methods are correlation and involve effect sizes involves the fear that the design may be viewed as being correlational. But the use of effect size estimates does not change the design of the study. ".any more than the use of an OVA technique as part of analysis makes the design experimental" (Thompson, 1987, p. 24).

The concept of effect size is not new, but only recently has its importance in reporting results in educational research has been recognized. And as Borg and Gall (1989, p. 364) suggest, "In sum, there is no simple answer to the problem of determining practical significance of research results. The effect size is just an aid to interpretation, albeit an important aid."

Summary

The use of statistical significance testing in educational research has often been misused. "The prevailing yes-no decision at the magic .05 level from a single research is a far cry from the use of informed judgement. Science simply doesn't work that way" (Cohen, 1990, p. 1311). Researchers need to recognize the limitations of significance testing and be wary of studies that base decisions of importance on such

testing. It is important to keep in mind that statistical significance does not imply that the results are necessarily meaningful or important. Effect size statistics aid in the interpretation of results and will provide a guide to the relative importance of the study.

References

Anastasia, A. (1988). Psychological testing (6th Ed.). New York: Macmillan Publishing Company.

Atkinson, D.R., Furlong, M.J., & Wampold, B.E. (1982). Statistical significance reviewer evaluation, and the scientific process: Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29, 189-194.

Borg, W.R. & Gall, M.D. (1989). Educational Research (5th Ed.). New York: Longman.

Chow, R.P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston, Inc.

Fisher, R.A. (1925). Statistical methods for research workers. Edinburgh: Oliver & Boyd.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

Hays, W.L. (1981). Statistics (3rd ed.). New York : Holt, Rinehart and Winston.

Holmes, C.B. (1990). The honest truth about lying with statistics. Springfield, Illinois: Charles C. Thomas Publisher.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Shaver, J.P. (1979). The productivity of educational research and the applied-basic distinction. Educational Researcher, 8, 3-9.

Tatsuoka, M.M. (1973). An examination of the statistical properties of a multivariate measure of strength of relationship. Urbana: University of Illinois. (ERIC Document Reproduction Service No. ED 099 406)

Thompson, B. (1987, April). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Education Research Association, Washington, DC. (ERIC Document Reproductive Service No. ED 287 868)

Thompson, B. (1988, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproductive Service No. ED 301 595)

Thompson, B. (1989a). Asking "What if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 22, 66-68.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Thompson, B. (1989c). The place of qualitative research in contemporary social science: The importance of post-paradigmatic thought. In B. Thompson (Ed.), Advances in Social Science Methodology, (Vol. 1, pp.1-42). Greenwich, CT: JAI Press Inc.

Webster's Ninth New Collegiate Dictionary. Springfield, MA: Merriam-Webster Inc.

Welge-Crow, P.A., LeCluyse, K.B. & Thompson, B. (1990, June). Looking beyond statistical significance: Result importance and result generalizability. Paper presented at the annual meeting of the American Psychological Society, Dallas, TX. (ERIC Document Reproductive Service No. ED forthcoming).

Table 1

Statistical Significance at Various Sample Sizes for a Fixed Effect Size and Two Levels of the Independent Variable

	A	B	C	D	E	F	G
1	SOURCE	SOS	R sq	df	MS	F calc	F Crit Dec
2							
3	SOSexp	*25*	*0.25*	2	12.50	3.17	3.52 Not Rej
4	SOSunexp	*75*		19	3.95		
5	SOStot	*100*		21	4.76		
6							
7	SOSexp	*25*	*0.25*	2	12.50	3.33	3.49 Not Rej
8	SOSunexp	*75*		20	3.75		
9	SOStot	*100*		22	4.55		
10							
11	SOSexp	*25*	*0.25*	2	12.50	3.50	3.47 Rej
12	SOSunexp	*75*		21	3.57		
13	SOStot	*100*		23	4.35		
14							
15	SOSexp	*25*	*0.25*	2	12.50	3.67	3.44 Rej
16	SOSunexp	*75*		22	3.41		
17	SOStot	*100*		24	4.17		

Note. As the sample size increases, tabled "critical F" values get smaller, error df gets larger, mean square error gets smaller and as a result "calculated F" also gets larger. For the purposes of these analyses numbers in italics remain fixed. From Thompson, (1989a), with permission.

Table 2

Statistical Significance at Various Sample Sizes for a Fixed Effect Size and Three Levels of the Independent Variable

	A	B	C	D	E	F	G
1	SOURCE	SOS	R sq	df	MS	F calc	F Crit Dec
2							
3	SOSexp	*25*	*0.25*	3	8.33	2.78	2.99 Not Rej
4	SOSunexp	*75*		25	3.00		
5	SOStot	*100*		28	3.57		
6							
7	SOSexp	*25*	*0.25*	3	8.33	2.89	2.98 Not Rej
8	SOSunexp	*75*		26	2.88		
9	SOStot	*100*		29	3.45		
10							
11	SOSexp	*25*	*0.25*	3	8.33	3.00	2.96 Rej
12	SOSunexp	*75*		27	2.78		
13	SOStot	*100*		30	3.33		
14							
15	SOSexp	*25*	*0.25*	3	8.33	3.11	2.95 Rej
16	SOSunexp	*75*		28	2.68		
17	SOStot	*100*		31	3.23		

Note. As the sample size increases, tabled "critical F" values get smaller, error df gets larger, mean square error gets smaller and as a result "calculated F" also gets larger. For the purposes of these analyses numbers in italics remain fixed. From Thompson, (1989a), with permission.